Stanford Department of Computer Science
Report No. STAN-CS-79-816

November 1979

AUTOMATING THE STUDY OF CLINICAL HYPOTHESES ON A TIME-ORIENTED DATA BASE THE RX PROJECT

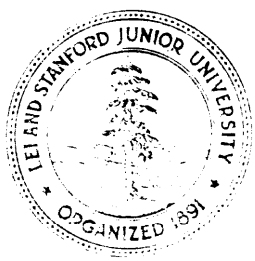
by

Robert L. Blum, M.D.

Research sponsored by

National Library of Medicine
and
National Center for Health Services Research

COMPUTER SCIENCE DEPARTMENT
Stanford University



AUTOMATING THE STUDY OF CLINICAL HYPOTHESES ON A TIME-ORIENTED DATA BASE: THE RX PROJECT

Robert L. Blum, M.D.

Stanford Medical Center Division of Clinical Pharmacology and
Stanford University Department of Computer Science
Margaret Jacks Hall, Stanford, California 94305

November, 1979

The RX Project

Robert L. Blum, M.D.
November, 1979

Abstract

The existence of large chronic disease data bases offers the possibility of studying hypotheses of major medical importance. An objective of the RX Project is to assist a clinical researcher with the tasks of experimental design and statistical analysis. A major component of RX is a knowledge base of medicine and statistics, organized as a frame-based, taxonomic tree. RX determines confounding variables, study design, and analytic techniques. It then gathers data, analyzes it, and interprets results. The American Rheumatism Association Medical Information System is used.

Automating the Study of Clinical Hypotheses on a Time-Oriented Data Base: The RX Project

Introduction

Computerized clinical data bases have great potential use for support of day to day health care delivery in hospitals and clinics, for monitoring the quality of health care delivery, and for assisting administrators with billing and resource management. They may also be used for performing clinical research, for assessing the efficacy of new diagnostic and therapeutic modalities, and for assisting with clinical management of individual patients. Over the past decade as the costs for computer equipment have plummeted, the number of data bases has greatly increased, and in the next decade we expect that clinical data bases will become ubiquitous.

One of the most important reasons for accumulating patient data on computers is the possibility of deriving medical knowledge from the stored observations. By knowledge - as distinct from data - we mean relationships between events or variables which are presumed to be true within a given setting. The following are examples of medical knowledge which one can imagine confirming on appropriate data bases: anti-hypertensive drugs prolong life-span for patients with severe hypertension, chloramphenicol may cause fat al aplastic anemia, certain arrythmias are strong predictors of sudden cardiac death. Similarly, one might imagine testing such important and controversial hypotheses as these: reduction in dietary cholesterol reduces atherogenesis, surgery for coronary artery disease increases life-span, or frequent administration of screening tests leads to decreased morbidity and mortality for certain cancers.

While the prospect of using clinical data bases to discover or to confirm medical hypotheses is tantalizing, there are many difficulties to be faced in using a body of data gathered in the past to adduce evidence on the validity of a hypothesis. If we are to use a data base to infer knowledge of clinical relationships, as distinct from its use as a mere repository for data, then we must confront all the problems of experimental design, adequacy of data, and confounding variables which complicate standard prospective clinical studies. If anything, the use of data gathered in the past typically for purposes other than clinical research, demands more stringent data analysis, study designs of greater sophisticat ion, and more thoughtful interpretation than does the use of prospectively collected data. [Blum:78]

The objective of the RX Project is to aid a clinical researcher in testing hypotheses of interest on a clinical data base by assisting with the task of experimental design and statistical analysis. While our major interest has been in testing the effects of drugs - hence, the name RX - the issues which we confront in this project are common to all systems which attempt to derive empirical knowledge from uncontrolled data.

The usual process of setting up a clinical study involves detailed medical knowledge of the drugs, diseases, and effects as well as considerable statistical expertise. The medical researcher, in collaboration with a statistician, may undertake a sequence of studies, using the results of initial studies to improve later designs.

We have tried to emulate parts of this process with the goal of improving the reliability of studies performed using clinical data bases. Our method involves the use of a knowledge base of relevant aspects of clinical medicine and statistical design in order to formulate an efficient experimental test of a researcher's hypothesis.

ARAMIS: A Time-Oriented Data Base

The data base which we use is the ARAMIS Data Base, the American Rheumatism Association Medical Information System, developed at Stanford University by Dr. James Fries and his research group. ARAMIS is implemented on a generalized data base system called TOD, Time-Oriented Data Base, developed by Prof. Gio Wiederhold and his associates. The TOD software is also used by the National Stroke Data Base Group, the Northern California Cancer Program, and the University of California Renal Transplant Program. The ARAMIS Data Base contains records of over 7,000 patients with a variety of rheumatologic disorders. [Fries:72][Weyl:75][Wiederhold:75,77]

Each patient's record consists of a matrix of values for a set of attributes which may be recorded each time the patient is seen in the clinic. The format of a hypothetical time-oriented record is illustrated below.

-VISIT NUMBER:	1	2	3
-DATE:	17 Jan 71	23 Jun 71	1 Jul 71
KNEE PAIN:	severe	mild	mild
FATIGUE:	moderate		moderate
TEMPERATURE:	38.5	37.5	36.9
DIAGNOSIS:	systemic lupus		
WHITE BLOOD COUNT:	3.5	4.7	4.3
CREATININE CLEARANCE:	45		65
BLOOD UREA NITROGEN:	36	33	
PREDNISONE (mgms/day)	30	25	20

Values for several hundred attributes can be recorded in ARAMIS. The attributes include signs, symptoms, lab tests, therapies, and indices of patient functional status [Hess:76]. TOD is implemented in PL/1; ARAMIS is stored on an IBM 370/3033 computer.

Overview of the RX Program

Conceptually, the major new component which RX adds to the TOD data base system is a knowledge base of clinical medicine and statistical study design, as well as the computational machinery for interpreting it. RX is implemented in INTERLISP, a dialect of LISP, a list-processing language highly suitable for knowledge engineering applications. The program is implemented on the DEC KI-10's of the SUMEX-AIM facility provided for the development of applications of artificial intelligence to medicine.

The categories of knowledge in the knowledge base arise naturally from a consideration of the steps involved in performing our overall goal - assisting a researcher with the task of testing a hypothesis on the data base; those steps are

1) Parse the hypothesis and determine the classification of variables in it.
2) Determine which other variables not in the original hypothesis might confound the relationship of primary interest.
3) Formulate an overall study design, and select analytic methods.
4) Create definitions for the events of relevance to the hypothesis.
5) Select appropriate patients in the data base.
6) Gather data by applying the event definitions to the selected records.
7) Run the data using the selected statistical method.
8) Interpret the results.
9) Based on these results, modify the study design and repeat the above sequence.
10) If results are conclusive, incorporate them into the knowledge base.

Several kinds of knowledge are involved in the many steps in this task, and we use a variety of formalisms for representing them. The main data structure of RX's knowledge base is a tree representing a taxonomy of relevant aspects of medicine and statistics. Each node in the tree is called a conceptual unit, or unit for short, and represents some object in the medical or statistical domain. For example, diagnostic categories such as endocrine or cardiac disorders are units standing for classes of disease; similarly, regression techniques and life-table methods are other units standing for classes of statistical methods.

This style of representation, usually called a frame-based knowledge representation, has been developed by a number of research groups [Bobrow:77][Stefik:78]. A frame or unit contains information describing properties of the object which it represents and its relationships with other units. Furthermore, it may contain procedures for performing a variety of tasks: recognizing instances of the object, causing other actions to be taken if certain conditions are met, and so on. These procedures may take the form of production rules [Davis:77], they may be calls to functions, or may actually generate functions which will be evaluated elsewhere. This style of computing is greatly facilitated by the LISP language.

As an example of a unit in RX's knowledge base, below we see parts of the unit representing HEPATITIS (unit names are all in capitals).

HEPATITIS

General: HEPATIC-DISORDERS
Special: (SEVERE-HEPATITIS CHRONIC-HEPATITIS)
Type: interval
Effects: CHOLESTEROL
Units of Measurement: international units of SGOT
Definition: SCOT > 70 international units
and TOTAL-BILIRUBIN > 2.0 mgms per deciliter
Reported-as: Maximum During Interval
Minimum-Duration: 14 days
Minimum-Observations: 3
Maximum-Gap: 7 days
Onset-Delay: 3 days
Carry-Over: 14 days

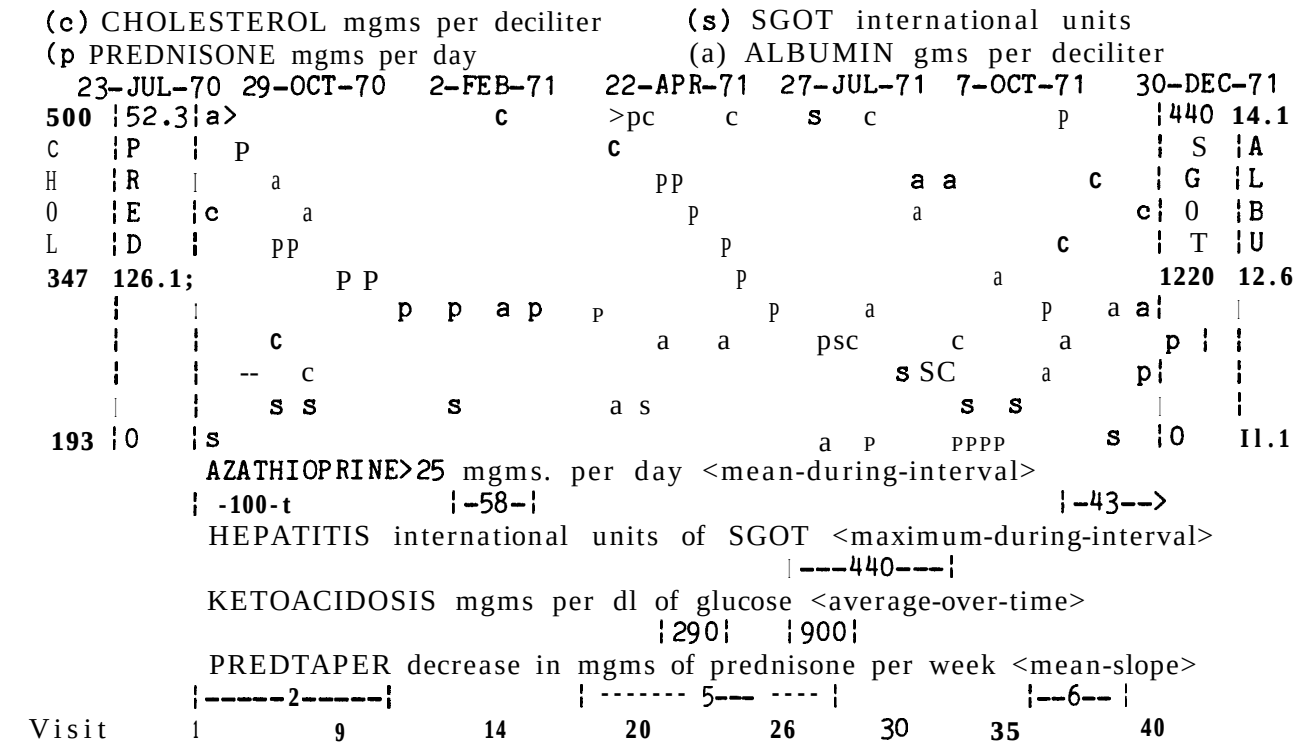
The properties, General and Special, indicate the parent and children nodes of this unit in the tree. "Type: interval" means that this node specifies an episode which has duration in time. On the Effects property we see that an alteration in serum cholesterol may occur although the exact relationship is not specified.

The properties Definition through Minimum-Observations all serve to specify the functional mapping between data base observations and this unit. In other words, they are used by the RX interpreter in deciding when the patient's record satisfied the definition of the unit, e.g. when the patient had hepatitis. Definitions may contain in them the names of other units which are not primary data base attributes. In the process of evaluating such definitions against a data base, the definitions for these other units are recursively evaluated in a depth-first invocation down to primary data base attributes.

The RX interpreter contains a collection of several dozen time-dependent predicates or functions for determining such things as days between events, adjacency of events, concurrency, various kinds of averages over time. One of the chief conceptual problems which we confront here is that of making inferences about episodes with duration in time based on point observations pertinent to these episodes as recorded in the data base.

As an illustration of our approach to inferring that an episode such as hepatitis occurred in a particular patient's history consider the HEPATITIS unit. We see from its properties that three observations with SGOT > 70 and TOTAL-BILIRUBIN > 2.0 must have been recorded within an interval of duration > 14 days but with a gap in observations not to exceed 7 days.

In the accompanying graph generated by RX we see certain events in an eighteen month period from one patient's record. The point values for cholesterol, prednisone, sgot (serum glutamic oxaloacetic transaminase), and albumin are simply a graphical display of values from ARAMIS. However, the episodes displayed below them - AZATHIOPRINE>25, HEPATITIS, KETOACIDOSIS, and PREDTAPER - are episodes which were abstracted from the data base data using the definitions in the appropriate units and time-dependent functions in the RX interpreter.



Using the Knowledge Base to Test a Specific Hypothesis

We will illustrate the method by which RX studies a clinical hypothesis by narrating its sequence of steps on one medically important hypothesis. Assume that the clinician has entered the following hypothesis.

H1: prednisone elevates cholesterol

The hypothesis H1 is parsed. H1 is of the form <unit1 relationship1 unit2>. Next we must determine what unitl, relationshipl, and unit2 are, that is, what their properties and relationships are to other conceptual units.

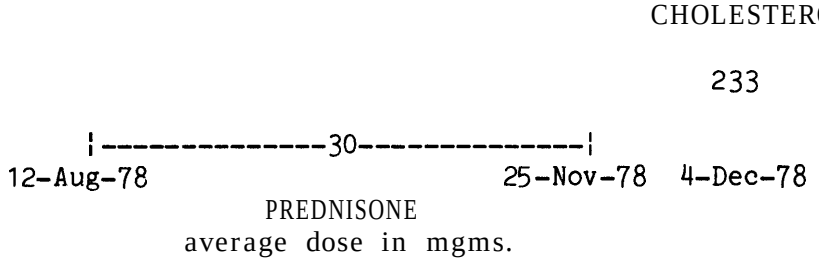
RX examines their ancestors in the tree of conceptual units in the knowledge base:

```
Class(PREDNISONE) =   STEROIDS, DRUGS, ACTIONS, ALL-UNITS
Class(ELEVATES) =     RELATIONSHIPS, ALL-UNITS
Class(CHOLESTEROL) =  CHEMISTRIES, LAB, STATES, ALL-UNITS
```

The function, class, simply follows parent links in the tree up to its root, ALL-UNITS. The meaning of this hierarchy is that of class membership. For example, PREDNISONE is one of the class, STEROIDS, which is one of the class, DRUGS, and so on.

By examining the properties of PREDNISONE and CHOLESTEROL, RX finds that PREDNISONE is given over some time interval and that CHOLESTEROL is a real-valued point in time. We also know that the event, measurement of CHOLESTEROL, must occur after the start of PREDNISONE. A production rule, stored under the unit STUDY-TYPES, fires narrowing the choice of STUDY-TYPE to one appropriate for the study of a drug given over an interval of time to a real-valued lab measurement determined atone point in time.

An instance satisfying this definition, so far, might look like this:



Other properties of PREDNISONE further restrict the space of possible event instances by adding constraints to the definition. The interval of prednisone must be greater than some minimum duration. The interval of time during which we wish to observe the cholesterol does not exactly coincide with the interval during which prednisone is administered. We would expect a delay in the onset of action of the drug and perhaps a carry-over in its effect beyond the period of administration. Since we are examining a new effect, we use default values, inherited from the class STEROIDS of which PREDNISONE is a member:
Onset-duration = 3 days, Minimum-duration = 30 days,
Carry-over = 1 day.

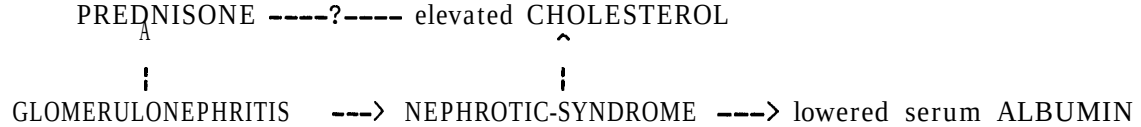
Of greater theoretical interest, knowledge stored under PREDNISONE and CHOLESTEROL, further specify the event definition by bringing to it the identity and nature of other influences: disqualifiers and covariates, the presence of which must be controlled for.

The disqualifiers are factors which, if present, will disqualify an event from further consideration in this study. Since PREDNISONE is one drug in the class STEROIDS, all other drugs in the class are disqualifiers for this study. That is, other STEROIDS, ACTH (steroid effect) and DEXAMETHASONE, must not have been given during the study interval. We find that anepisode of HEPATITIS is a disqualifier inferred from information under CHOLESTEROL as is an episode of KETOACIDOSIS. Note that these disqualifying events must also be fully specified in terms of their duration and timing with respect to the prednisone interval and the cholesterol measurement.

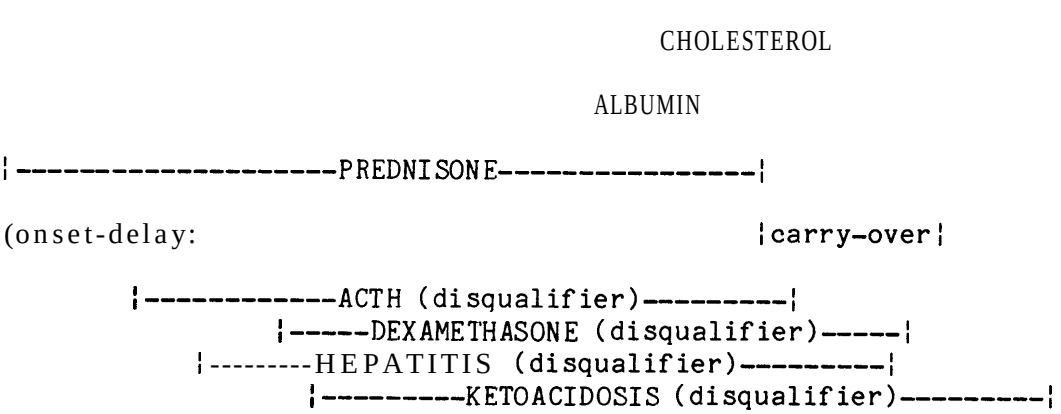
The covariates are those factors whose presence must be controlled for since they are known by the knowledge base to influence the units in the hypothesis. Covariates are collected by applying productions and examining lists associated with PREDNISONE and CHOLESTEROL.

The most worrisome class of covariates are those factors which caused the drug to be given in the first place. In the context of this example, the question is, "if prednisone is positively correlated with cholesterol, is it a direct effect or is it due to other factors." On the Indications list for steroids we find GLOMERULONEPHRITIS, which in turn has as a manifestation, NEPHROTIC-SYNDROME. But elevated CHOLESTEROL is a manifestation of NEPHROTIC-SYNDROME which we assess by its depression of serum ALBUMIN.

The chain of possible causality now looks like this:



and the definition for a study event looks like this:



After preliminary definitions for relevant variables have been collected, a suitable statistical method is selected using knowledge stored in the form of production rules, located under units corresponding to the various statistical methods in RX's repertoire. A step-wise, hierarchical regression design was chosen for evaluation of the current hypothesis. This choice was based on the number of variables in the model, the value types of the variables, their presumed causal relationship, and their durations over time.

The initial model which has been created using information gathered by the RX interpreter from the units corresponding to the variables of the hypothesis, their inferred covariates and disqualifiers, and from the unit for multiple regression is

cholesterol = b1*albumin + b2*previous-cholesterol + b3*prednisone + constant

No data are to be taken at those times when acth, dexamethasone, hepatitis, or ketoacidosis might be influencing the cholesterol values. A lag-term, previous-cholesterol, is added to assess the effect of autocorrelation. Further terms are also included to adjust for interperson variability in the estimate of residual error.

A study editor displays the entire study design to the user, permitting him to edit certain parts of it. For example, a user may decide that he wishes to include values of cholesterol which occurred during intervals of hepatitis, but that he would like to exclude values taken during all hospitalizations, adding that to the list of disqualifiers.

Following the editing session, the interpreter constructs a set of functions containing time-dependent predicates which when applied to the data base will return sets of values satisfying all the defined constraints.

Next a set of patient records must be chosen on which to apply the constructed definitions for the study event. This task itself may be approached by techniques of varying sophistication. Our current heuristic simply ranks the patients according to quantity of relevant data in their records and selects a sufficient number to allow reasonable statistical power. Data is then gathered by application of the study event definitions to the selected records.

The extracted values must then be analyzed statistically. For this purpose we use SPSS, the Statistical Package for the Social Sciences. [Nie:75] One of the RX programs creates an SPSS "source deck" containing card images of the appropriate commands along with the extracted sets of values. RX then calls the operating system and runs SPSS on the source file, which creates a file containing the listing of its results. Program control then returns to RX where important results from the listing are extracted and interpreted.

For the hypothesis of this example, the linear model was found to account for 53% of the variation in cholesterol with a highly significant F value. The coefficients for all terms of the model, including prednisone, were also highly significant.

Limitations

We must emphasize the limitations which constrain any system that attempts to draw conclusions based on past observations. Most obviously, the strength of conclusions depends on the quantity and quality of data. Statistical estimates of confidence intervals may allow us to assess the adequacy of the amount of data available, but the completeness of the model is an extra-statistical issue. In the example above, it may be that phenomena external to the model are causing the increase in cholesterol which we have attributed to prednisone. If these phenomena are reflected by values of other attributes in the data base, then there is a possibility for finding and controlling for them. Otherwise, we have no means for controlling for their influence.

Current Status and Conclusions

At this time - November, 1979 - the RX knowledge base contains about 200 units; however, only 75 are completely filled in. There are 40 time-dependent functions used to map from the data base values onto the defined units. Although the current system contains knowledge on 12 statistical methods, we have only programmed interfaces to SPSS for simple, partial, and multiple regression. The system which is currently implemented runs hypotheses like the prednisone/cholesterol example using about 10 cpu minutes on a DEC KI-10. Several other hypotheses of medical interest have been tested.

The task of studying hypotheses using a time-oriented data base is complex and requires detailed knowledge of clinical medicine, experimental design, and statistics. The RX Program assists a clinical researcher in this task by designing a study of his hypothesis including definitions for complex events. It gathers the appropriate data, analyzes it, and interprets the results. Future work will address methods for extending our capabilites for analyzing residuals and revising regression models. We are also currently exploring means for discovering hypotheses of interest in the data base.

Acknowledgements

Sincere thanks to other members of the RX Project including Jan Clayton, Jerrold Kaplan, Patricia Pickett , Gio Wiederhold ; members of the ARAMIS Project particularly Alison Harlow and Guy Kraines; and members of the Stanford Heuristic Programming Project particularly Lawrence Fagan, Peter Friedland, Mark Stefik, and William VanMelle.

Funding for this project was provided by the National Center for Health Services Research and by NIH Grant LM 03370 from the National Library of Medicine. The author is the recipient of a Post-Doctoral Research Fellowship from the Pharmaceutical Manufacturers Association Foundation. SUMEX-AIM is supported by the Biotechnology Resources Program through NIH Grant RR-00785.

References

Blum, Robert L. and Wiederhold, Gio: Inferring Knowledge from Clinical Data Banks Utilizing Techniques from Artificial Intelligence. "Proc. 2nd Annual Symp. on Comp. Applic. in Med. Care," pp. 303-307, IEEE, Washington, D.C., November 5-9, 1978

Bobrow, Daniel G.; Winograd, Terry: An Overview of KRL, A Knowledge Representation Language. "Cognitive Science" vol. 1, no. 1, 1977

Davis, Randall B.; Buchanan, Bruce G.; Shortliffe, Edward H.: Production Rules as a Representation for a Knowledge-Based Consultation Program. "Artificial Intelligence" 8:15-45, 1977

Fries, James F.: Time-Oriented Patient Records and a Computer Databank. "Jour. of the Amer. Med. Assoc." 222:12:1536, December 18, 1972

Hess, Evelyn V.: A Uniform Database for Rheumatic Diseases. "Arthritis and Rheumatism," 19:3, pp. 645-648, May-June, 1976

Nie, Hull, Jenkins, Steinbrenner, Bent: Statistical Package for the Social Sciences. 2nd Edition. McGraw-Hill, 1975

Stefik, Mark: An Examination of a Frame-Structured Representation System. Stanford Heuristic Programming Project Memo HPP-78-13, Sept., 1978

Weyl, Stephen; Fries, J.; Wiederhold, G.; Germano, F.: A Modular Self-Describing Clinical Databank System. "Comp. and Biomed. Res.," 8:3, pp. 279-293, June, 1975

Wiederhold, Gio; Fries, James F.: Structured Organization of Clinical Data Bases. "AFIPS Conference Proceedings" 44: 479-485, 1975

Wiederhold, Gio: Database Design. McGraw-Hill, 1977